D174–D181 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh128

AANT: the Amino Acid–Nucleotide Interaction Database

Michael M. Hoffman[1,2], Maksim A. Khrapov[2], J. Colin Cox[1], Jianchao Yao[1], Lingnan Tong[1,3] and Andrew D. Ellington[1,2,*]

[1]Institute for Cellular and Molecular Biology, [2]Department of Chemistry and Biochemistry and [3]Department of Computer Sciences, University of Texas at Austin, Austin, TX 78712-0159, USA

Received August 15, 2003; Revised October 7, 2003; Accepted October 24, 2003

ABSTRACT

We have created an Amino Acid–Nucleotide Interaction Database (AANT; http://aant.icmb.utexas.edu/) that categorizes all amino acid–nucleotide interactions from experimentally determined protein–nucleic acid structures, and provides users with a graphic interface for visualizing these interactions in aggregate. AANT accomplishes this by extracting individual amino acid–nucleotide interactions from structures in the Protein Data Bank, combining and superimposing these interactions into multiple structure files (e.g. 20 amino acids × 5 nucleotides) and grouping structurally similar interactions into more readily identifiable clusters. Using the Chime web browser plug-in, users can view 3D representations of the superimpositions and clusters. The unique collection and representation of data on amino acid–nucleotide interactions facilitates understanding the specificity of protein–nucleic acid interactions at a more fundamental level, and allows comparison of otherwise extremely disparate sets of structures. Moreover, by modularly representing the fundamental interactions that govern binding specificity it may prove possible to better engineer nucleic acid binding proteins.

INTRODUCTION

Protein–nucleic acid interactions have been variously described by focusing on either the protein's primary sequence (such as arginine-rich motifs) (1), the protein's tertiary structure (e.g., helix–turn–helix motifs) (2,3), the nature of the nucleic acids that are bound (double-stranded versus single-stranded nucleic acids) (4) or the conformational changes that occur during complex formation ('induced-fit' versus 'lock-and-key' models) (5). However, the data that underlies all of these approaches are the discrete interactions between amino acids and nucleotides. While such amino acid–nucleotide interactions are structurally diverse, it is nonetheless clear that in at least some instances they fall into recognizable classes, such as the 'arginine fork' motif in which arginine forms a pseudo-Hoogsteen pairing with guanosine (6).

The utility of examining amino acid–nucleotide interactions has also been noted by other researchers, who have previously established databases such as the Protein–Nucleic Acid Interaction server (http://www.biochem.ucl.ac.uk/bsm/DNA/server/) (7,8) and the Protein–Side Chain Interactions database (http://www.biochem.ucl.ac.uk/bsm/sidechains/) (9) created by the Thornton group at University College London. However, these databases were created in order to study non-overlapping structures, and hence contain a relatively small number of structures. In addition, they do not contain interactions involving the peptide backbone, sugar or phosphate backbone. Finally, these databases have focused exclusively on DNA structures, and contain no information regarding protein–RNA interactions. While Treger and Westhof (10) have previously published an excellent analysis of a similarly limited number of protein–RNA interactions, there is no available database associated with this analysis.

We have therefore developed a comprehensive amino acid–nucleotide interaction database (AANT; http://aant.icmb.utexas.edu/) that deconstructs the structures of all known protein–nucleic acid interactions into sets of amino acid–nucleotide interactions. This database should prove extremely useful in determining the extent and breadth of amino acid–nucleotide interactions, in intelligently categorizing such interactions, and eventually in using preferred appositions in the design of altered or novel protein–nucleic acid interactions.

METHODS

Deriving interaction super-models from experimental structures

The software behind AANT consists of a series of Perl and Python scripts, which automatically update AANT once per week. AANT searches the Protein Data Bank (11) for experimental structures that contain both a protein and either a DNA or RNA molecule, and downloads any structures newer than the latest PDB CD-ROM set (Release 104, April 2003). In the case of structures generated with NMR data, the

*To whom correspondence should be addressed. Tel: +1 512 232 3424; Fax: +1 512 471 7014; Email: andy.ellington@mail.utexas.edu
Present address:
Michael M. Hoffman, EMBL—European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

software only considers the first alternative structure. AANT then uses the program HBPLUS (12) to predict hydrogen bonds between single nucleotide residues and single amino acid residues. The bonded structures are broken into scores of individual interactions between the base, the sugar or the phosphate of a nucleotide residue and the side chain or peptide backbone of an amino acid residue.

AANT creates sub-models for each individual amino acid–nucleotide interaction containing only the residues involved in the interaction. If an amino acid residue or nucleotide residue participates in multiple interactions, it will be duplicated in multiple sub-models. AANT transforms each sub-model into a new coordinate system, centering on a fixed atom in the nucleotide moiety involved (base, sugar or phosphate), while keeping the internal geometry of the interaction intact (13). AANT then assigns each sub-model to an interaction class, defined by the attributes of the interaction involved: the nucleotide (five possibilities), the nucleotide moiety (base, sugar or phosphate; three possibilities), the amino acid (20 possibilities) and the amino acid moiety (side chain or peptide backbone; two possibilities). While there are therefore 600 (= 5 × 3 × 20 × 2) potential interaction classes, only 423 of these classes have actually been observed in experimentally determined structures as of this writing. After creating and classifying all the sub-models, AANT creates a super-model for each interaction class that is a superposition of each sub-model. For a given interaction class, the super-model contains closely overlapping nucleotides surrounded by a more dispersed constellation of amino acids, approaching the nucleotides from all of the directions and with all of the geometries observed in natural structures.

As an example, consider an experimental structure of a protein–DNA complex. AANT predicts dozens of hydrogen bonds between the protein and DNA, including a hydrogen bond between a glutamine side chain and an adenine nucleobase. For the purpose of analyzing this interaction, AANT segregates these two residues from the rest of the structure into a sub-model and for the moment ignores the other amino acid and nucleotide residues. While the adenosine residue in question might be located anywhere within the original structure, AANT defines a new coordinate system for the sub-model so that nitrogen 7 is located at (0, 0, 0), carbon 6 is located at (0, 0, z) and nitrogen 3 is located at (x, 0, z'), where x, z and z' are the three quantities that will conserve the distances and angles between the atoms from the original structure. AANT transforms all atoms in the adenosine and glutamine residues to this new coordinate system, conserving the original distances and angles throughout. AANT then creates a super-model, superimposing this transformed sub-model with the transformed sub-models of all other interactions between glutamine side chains and adenine nucleobases, including sub-models derived from other experimental structures. Since all of these sub-models have undergone the same transformation, the nitrogen 7 atoms of the adenosines are superimposed exactly on top of each other, the rest of the adenosine atoms will be reasonably close to their analogous atom from another residue, and amino acid residues will be scattered around the periphery of the superimposed adenosines, in accord with their original distances and angles from their cognate adenosines. Sugar and phosphate moieties of the adenosine nucleotide residues will also be scattered to a certain degree, depending on their relative conformations with respect to their nucleobases and nitrogen 7 atoms. The output from AANT for this particular example is displayed in Figure 1.

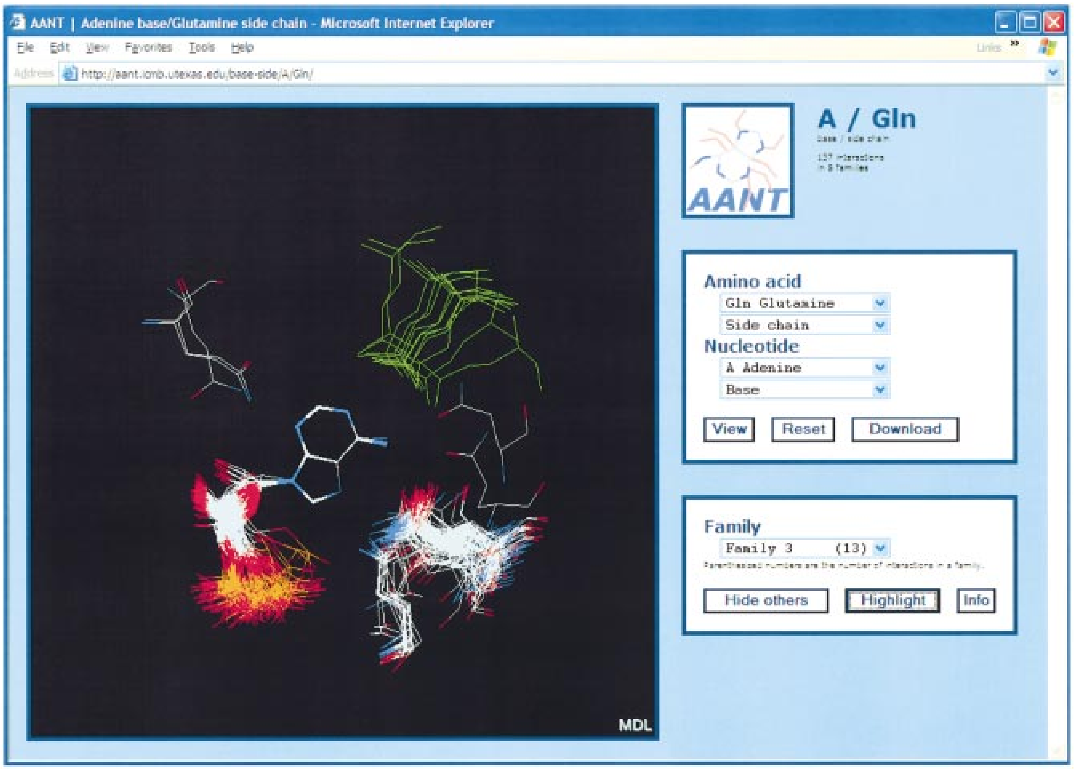
Classifying interactions into clusters

After AANT has created super-models containing all known amino acid–nucleotide interactions, it uses an algorithm to group the amino acid residues within a super-model into clusters based on the 'simple cluster-seeking algorithm' described by Tou and Gonzalez (14). For clustering purposes, AANT defines a distance score between two amino acid residues as the number of non-hydrogen atoms in one residue multiplied by the square of the root-mean-square distance between the two residues. AANT begins the clustering process by assigning an initial interaction to a cluster, and then assigns to that cluster all other interactions that fall within a given distance score (initially set at 50 Å^2) of the initial interaction or of any interaction that has been added to the initial cluster. AANT repeats this process until it cannot assign any more members to this first cluster, and then begins a second cluster, iterating until it has assigned all interactions to a cluster. If the clustering algorithm produces more clusters than 13, AANT increases the distance score threshold slightly and iterates the process until 13 or fewer clusters result. (We limit the number of clusters to 13 because two PDB 'chains' are assigned to each cluster, and PDB files typically do not contain more than 26 chains.)

For example, consider a series of superimposed glutamine–adenosine interactions found in the same super-model. If the first glutamine residue has a distance score of 60 Å^2 from the second, but both have a distance score of just 30 Å^2 from the third residue, they will all be assigned to the same cluster. If all three of these residues have distance scores that are >50 Å^2 away from all other glutamine residues in the super-model, then AANT will assign the other residues into new clusters. If this process yields more than 13 clusters, then AANT will discard the clustering assignments and begin the process anew using a distance score threshold of 54 Å^2.

DISCUSSION

As of this writing, the PDB contained 930 solved structures of complexes between proteins and nucleic acid molecules. The AANT database extracts all interactions between single bases and single amino acid residues from these structures. For each of these classes of interactions, the AANT software generates a 3D superimposition and clusters structurally similar interactions into families based on spatial similarity. For example, AANT classifies the 194 known interactions between an adenine base and an arginine side chain into 12 families. Using the free Chime chemical display browser plug-in (http://www.mdlchime.com/), the 3D superimposition and renderings of families can be visualized and manipulated in real time. Researchers may also download structural models for further analysis using any of the publicly available tools for manipulating PDB structures, such as RasMol (15) and DeepView (16). However, by using the Chime plug-in it should be possible for users to quickly segue between different structural representations and to thereby generate their own structural hypotheses. In addition to the 3D renderings, AANT


AANT | Adenine base/Glutamine side chain - Microsoft Internet Explorer

Address http://aant.icmb.utexas.edu/base-side/A/Gln/

A / Gln

base / side chain

137 interactions
in 8 families

Amino acid: Gln Glutamine; Side chain

Nucleotide: A Adenine; Base

Family: Family 3 (13)

Parenthesized numbers are the number of interactions in a family.

Figure 1. Screenshot of AANT. Some 137 superimposed interactions between glutamine side chains and adenine nucleobases are shown. The prevalence of different kinds of interactions is readily apparent. AANT has grouped these interactions into eight clusters. Users may highlight any cluster or hide unwanted clusters. The amino acid residues from Family 3 have been highlighted in green. The superimposed amino acid and nucleotide residues from other families are shown using the Corey–Pauling–Kultun (CPK) color scheme. Users may also rotate the superimposed model in three dimensions, display the partners using space-filling models and save the structure locally.

also provides a novel, simplified 2D schematic that shows all of the predicted interactions between a given protein and a complexed nucleic acid (Fig. 2).

While there is no code *per se* for amino acid–nucleotide interactions, some combinations of amino acids and nucleotides are clearly preferred in the context of protein–nucleic acid interactions. There have been several previous attempts (9,10,17–20) to count amino acid–nucleotide interactions and to use these data to draw conclusions about the specificity of protein–nucleic acid interactions. However, these previous attempts either did not cluster interactions into structural subsets, or involved only a relatively limited number of structures.

For example, the Nucleic Acid Interaction Library (NAIL) (20) contains theoretical predictions (as opposed to experimental models) of several different kinds of interactions, and overlaps with AANT only in terms of nucleobase–amino acid side chain interactions. When experimental structures were used to evaluate NAIL's completeness, the authors observed no classes of interactions that were not already within NAIL. This contrasts strongly with our own observations, as AANT contains more classes of experimentally observed amino acid–nucleobase interactions than are included in NAIL. We conjecture that the AANT software observed more types of nucleobase–side chain interactions primarily because it uses different hydrogen-bond prediction software.

In addition, Thornton and her co-workers have previously generated the Protein–Side Chain Interactions database and have analyzed the apposition of amino acids and nucleotides from 129 non-homologous protein–DNA complexes (9). Similarly, Treger and Westhof (10) analyzed the apposition of amino acids and nucleotides from 45 non-homologous protein–RNA complexes. In contrast, AANT has, as of this writing, almost an order of magnitude more structures, and is continually updated. As an example of the difference in coverage, the Protein–Side Chain Interaction database reports 56 hydrogen-bond interactions with arginine; AANT reports and can classify 100 times that many interactions (>3300 for Arg–DNA, >2700 for Arg–RNA).

Overall, the larger AANT database largely validates the findings of the smaller statistical samplings, but also provides additional data for analysis and discussion. For example, in a study of 45 non-homologous protein–RNA interactions, Treger and Westhof (10, their table 7) found that 20% of hydrogen-bonded or ionic interactions were between amino acids and ribose, 43% were with phosphate and 38% were with bases. A similar study by Jones and co-workers (8, their table 5) encompassed only 32 non-homologous protein–RNA interactions and found that the proportions were ribose 16%, phosphate 34%, bases 50%. In the AANT database the proportions are ribose 23%, phosphate 51% and bases 26% (Table 1). The numerical discrepancies between these studies may be due to the sample size or to the ways in which hydrogen-bonded contacts were found or counted, but it seems that phosphate contacts are greatly preferred relative to contacts with ribose, typically by a 2:1 margin. This skewing is further exacerbated in protein–DNA interactions, where only 2% of the contacts are with deoxyribose. Luscombe and co-workers previously made this observation (9, their table 2) based on a much smaller data set.

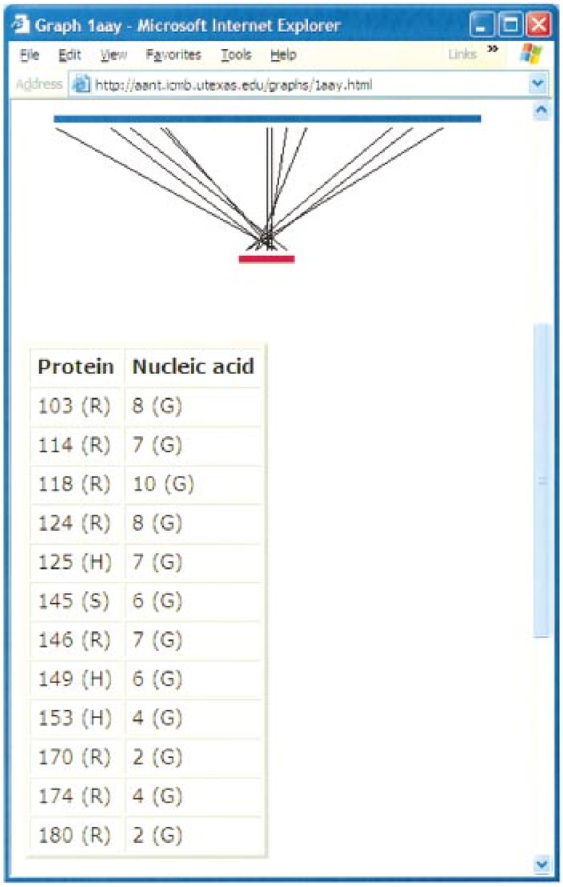
The data in Table 1 also show that while researchers frequently focus on interactions between amino acid side chains and nucleobases (13,20,21), these only make up 19.8% of the total, predicted hydrogen-bond interactions found in the database. Individual data for DNA (21.7%) and RNA (17.2%) are found in Table 1. These statistics reinforce the conclusion that the peptide backbone and the nucleotide sugar and phosphate contribute significantly to the affinity and specificity of protein–nucleic acid interactions.

Statistical breakdowns of the identities of amino acid–nucleotide interactions are presented in Tables 2–5 and graphically in Figure 3. The χ^2 values (not including Yate's correction, because of the generally large sample sizes) were calculated for individual amino acid–nucleotide interactions. Each χ^2 value corresponds to a 2 × 2 contingency table

Table 1. Interactions between substructures

	Peptide backbone (%)	Side chain (%)	Total (%)
DNA			
base	4.1	21.7	25.8
phosphate	25.6	46.6	72.2
sugar	0.3	1.7	2.0
total	30.0	70.0	
RNA			
base	9.2	17.2	26.4
phosphate	10.3	40.4	50.7
sugar	8.5	14.4	22.9
total	28.0	72.0	

The number of interactions between nucleotide moieties (vertical) and amino acid moieties (horizontal) are given as a fraction of the total number of interactions in AANT with DNA and RNA.



Graph 1aay - Microsoft Internet Explorer

http://aant.icmb.utexas.edu/graphs/1aay.html

Protein	Nucleic acid
103 (R)	8 (G)
114 (R)	7 (G)
118 (R)	10 (G)
124 (R)	8 (G)
125 (H)	7 (G)
145 (S)	6 (G)
146 (R)	7 (G)
149 (H)	6 (G)
153 (H)	4 (G)
170 (R)	2 (G)
174 (R)	4 (G)
180 (R)	2 (G)

Figure 2. Example of a 2D rendering of protein–nucleic acid interactions. The Zif268 zinc-finger peptide (blue) and one strand (red) from a duplex 11-mer oligonucleotide that binds to the peptide (PDB ID: 1aay) are laid out as lines. Moving from left to right, the cross lines join amino acids from N-terminus to C-terminus with nucleotides from 5′ to 3′. AANT generates a similar diagram for each experimental structure of a protein–nucleic acid complex.

Table 2. Interactions between amino acids and DNA nucleotides (% in parentheses)

	Adenosine	Cytidine	Guanosine	Thymidine	Total
Alanine	115 (28.1)	80 (19.6)	75 (18.3)	139 (34.0)	409
Arginine	700 (21.0)	570 (17.1)	1333 (39.9)	734 (22.0)	3337
Asparagine	342 (30.6)	201 (18.0)	260 (23.3)	315 (28.2)	1118
Aspartate	50 (20.1)	111 (44.6)	79 (31.7)	9 (3.6)	249
Cysteine	13 (18.8)	20 (29.0)	9 (13.0)	27 (39.1)	69
Glutamine	256 (37.9)	106 (15.7)	152 (22.5)	161 (23.9)	675
Glutamate	10 (4.5)	138 (62.7)	40 (18.2)	32 (14.5)	220
Glycine	126 (19.4)	130 (20.0)	203 (31.2)	191 (29.4)	650
Histidine	120 (26.9)	41 (9.2)	181 (40.6)	104 (23.3)	446
Isoleucine	37 (33.0)	11 (9.8)	39 (34.8)	25 (22.3)	112
Leucine	4 (3.9)	25 (24.3)	31 (30.1)	43 (41.7)	103
Lysine	512 (23.2)	404 (18.3)	635 (28.8)	657 (29.8)	2208
Methionine	13 (25.0)	14 (26.9)	13 (25.0)	12 (23.1)	52
Phenylalanine	26 (27.7)	25 (26.6)	15 (16.0)	28 (29.8)	94
Proline	2 (25.0)	2 (25.0)	2 (25.0)	2 (25.0)	8
Serine	440 (28.0)	219 (14.0)	424 (27.0)	486 (31.0)	1569
Threonine	327 (20.4)	332 (20.7)	395 (24.6)	549 (34.2)	1603
Tryptophan	18 (14.0)	38 (29.5)	29 (22.5)	44 (34.1)	129
Tyrosine	159 (22.9)	180 (26.0)	183 (26.4)	171 (24.7)	693
Valine	61 (28.8)	50 (23.6)	53 (25.0)	48 (22.6)	212
Total	3331 (23.9)	2697 (19.3)	4151 (29.7)	3777 (27.1)	13956

Percentages in parentheses indicate the fraction of total interactions between a particular amino acid and one of five nucleotides. The information in this table is also summarized in Figure 3.

Table 3. Interactions between amino acids and RNA nucleotides (% in parentheses)

	Adenosine	Cytidine	Guanosine	Uridine	Total
Alanine	22 (15.0)	14 (9.5)	60 (40.8)	51 (34.7)	147
Arginine	500 (18.4)	902 (33.2)	821 (30.2)	493 (18.2)	2716
Asparagine	114 (17.4)	157 (23.9)	158 (24.0)	228 (34.7)	657
Aspartate	31 (8.6)	98 (27.1)	149 (41.3)	83 (23.0)	361
Cysteine	21 (45.7)	0 (0.0)	20 (43.5)	5 (10.9)	46
Glutamine	71 (14.2)	103 (20.6)	197 (39.3)	130 (25.9)	501
Glutamate	118 (26.5)	91 (20.4)	181 (40.6)	56 (12.6)	446
Glycine	99 (21.6)	125 (27.3)	156 (34.1)	78 (17.0)	458
Histidine	62 (23.0)	74 (27.5)	89 (33.1)	44 (16.4)	269
Isoleucine	7 (7.4)	43 (45.3)	25 (26.3)	20 (21.1)	95
Leucine	16 (13.6)	44 (37.3)	36 (30.5)	22 (18.6)	118
Lysine	444 (24.5)	520 (28.7)	595 (32.8)	254 (14.0)	1813
Methionine	14 (20.0)	20 (28.6)	14 (20.0)	22 (31.4)	70
Phenylalanine	2 (3.1)	13 (20.3)	49 (76.6)	0 (0.0)	64
Proline	35 (38.0)	9 (9.8)	19 (20.7)	29 (31.5)	92
Serine	247 (28.1)	290 (33.0)	245 (27.9)	97 (11.0)	879
Threonine	251 (33.3)	149 (19.8)	264 (35.1)	89 (11.8)	753
Tryptophan	11 (15.1)	23 (31.5)	21 (28.8)	18 (24.7)	73
Tyrosine	162 (36.5)	136 (30.6)	82 (18.5)	64 (14.4)	444
Valine	28 (19.4)	76 (52.8)	17 (11.8)	23 (16.0)	144
Total	2255 (22.2)	2887 (28.5)	3198 (31.5)	1806 (17.8)	10146

Percentages in parentheses indicate the fraction of total interactions between a particular amino acid and one of five nucleotides. The information in this table is also summarized in Figure 3.

Table 4. Statistical data for protein–DNA interactions in AANT

	Adenosine			Cytidine			Guanosine			Thymidine			Total
Amino acid	Observed	Expected	χ^2	Observed	Expected	χ^2	Observed	Expected	χ^2	Observed	Expected	χ^2	
Alanine	115	98	4	80	79	0	75	122	26	139	111	10	409
Arginine	700	796	20	570	645	14	1333	993	218	734	903	57	3337
Asparagine	342	267	30	201	216	1	260	333	24	315	303	1	1118
Aspartate	50	59	2	111	48	104	79	74	0	9	67	71	249
Cysteine	13	16	1	20	13	4	9	21	9	27	19	5	69
Glutamine	256	161	77	106	130	6	152	201	18	161	183	4	675
Glutamate	10	53	46	138	43	270	40	65	14	32	60	18	220
Glycine	126	155	8	130	126	0	203	193	1	191	176	2	650
Histidine	120	106	2	41	86	30	181	133	26	104	121	3	446
Isoleucine	37	27	5	11	22	7	39	33	1	25	30	1	112
Leucine	4	25	23	25	20	2	31	31	0	43	28	11	103
Lysine	512	527	1	404	427	2	635	657	1	657	598	10	2208
Methionine	13	12	0	14	10	2	13	15	1	12	14	0	52
Phenylalanine	26	22	1	25	18	3	15	28	9	28	25	0	94
Proline	2	2	0	2	2	0	2	2	0	2	2	0	8
Serine	440	374	17	219	303	33	424	467	6	486	425	14	1569
Threonine	327	383	12	332	310	2	395	477	23	549	434	47	1603
Tryptophan	18	31	7	38	25	9	29	38	3	44	35	3	129
Tyrosine	159	165	0	180	134	21	183	206	4	171	188	2	693
Valine	61	51	3	50	41	3	53	63	2	48	57	2	212
	3331	3331		2697	2697		4151	4151		3777	3777		13956

These values cover all classes of interactions between the amino acid (side chain and backbone) and the nucleotide (sugar, phosphate and base). The χ^2 values were calculated for 2 × 2 contingency tables, as described in the text.

representing the association of a given amino acid with a given nucleotide, e.g. the four cells in the 2 × 2 table associated with alanine–deoxyadenosine interactions would be the value actually shown in Table 4 (alanine–deoxyadenosine, 115), alanine–non-deoxyadenosine (294), non-alanine–deoxyadenosine (3216) and non-alanine–non-deoxyadenosine (10 331). Hence, all χ^2 values represent probabilities with one degree of freedom. This analysis is similar to the one performed by Treger and Westhof (10). In general, the positively charged amino acids lysine and arginine mediate the largest number of contacts in protein–nucleic acid interactions, while cysteine and non-polar amino acids mediate the fewest. Conversely, interactions with guanine are overrepresented within almost all of the amino acid classes. These conclusions are again roughly similar to those that have been drawn in previous studies (9,10), but now more detailed analyses of the preferences of amino acids for nucleotides can be carried out on a much larger data set. For example, Treger and Westhof (10) have suggested

Table 5. Statistical data for protein–RNA interactions in AANT

	Adenosine			Cytidine			Guanosine			Uridine			Total
Amino acid	Observed	Expected	χ^2	Observed	Expected	χ^2	Observed	Expected	χ^2	Observed	Expected	χ^2	
Alanine	22	33	5	14	42	26	60	46	6	51	26	29	147
Arginine	500	604	31	902	773	41	821	856	3	493	483	34	2716
Asparagine	114	146	10	157	187	7	158	207	18	228	117	103	657
Aspartate	31	80	40	98	103	0	149	114	17	83	64	6	361
Cysteine	21	10	15	0	13	18	20	14	3	5	8	1	46
Glutamine	71	111	20	103	143	16	197	158	15	130	89	19	501
Glutamate	118	99	5	91	127	15	181	141	18	56	79	9	446
Glycine	99	102	0	125	130	0	156	144	1	78	82	0	458
Histidine	62	60	0	74	77	0	89	85	0	44	48	0	269
Isoleucine	7	21	12	43	27	13	25	30	1	20	17	1	95
Leucine	16	26	5	44	34	5	36	37	0	22	21	0	118
Lysine	444	403	7	520	516	0	595	571	2	254	323	47	1813
Methionine	14	16	0	20	20	0	14	22	4	22	12	9	70
Phenylalanine	2	14	14	13	18	2	49	20	61	0	11	13	64
Proline	35	20	13	9	26	16	19	29	5	29	16	12	92
Serine	247	195	19	290	250	10	245	277	6	97	156	34	879
Threonine	251	167	58	149	214	30	264	237	5	89	134	22	753
Tryptophan	11	16	2	23	21	0	21	23	0	18	13	3	73
Tyrosine	162	99	55	136	126	1	82	140	37	64	79	4	444
Valine	28	32	1	76	41	42	17	45	26	23	26	0	144
	2255	2255		2887	2887		3198	3198		1806	1806		10146

These values cover all classes of interactions between the amino acid (side chain and backbone) and the nucleotide (sugar, phosphate and base). The χ^2 values were calculated for 2 × 2 contingency tables, as described in the text.

that in protein–RNA interactions asparagine prefers to hydrogen-bond with uridine ($p < 0.0001$, one degree of freedom), and that serine prefers adenosine ($p < 0.0001$), findings that are confirmed in the larger database. Similarly, Luscombe and co-workers (9) found that in protein–DNA interactions arginine and lysine prefer to hydrogen-bond with guanosine. We strongly confirm the preference of arginine for guanosine, but our larger data set finds no statistical support for a preference of lysine for guanosine ($p < 0.27$); if anything, lysine may prefer thymidine ($p < 0.002$). In making these comparisons, one caveat is that AANT quantifies interactions between protein backbone residues and nucleotides, while other studies typically do not. The inclusion of protein backbone residues and the larger sample size in AANT allows other hydrogen-bonding preferences not previously noted to become apparent. For example, in protein–RNA complexes we now note the preference of both threonine and tyrosine for adenosine ($p < 0.0001$), and the preference of threonine for thymidine ($p < 0.0001$). In protein–DNA complexes glutamine is revealed to prefer adenosine ($p < 0.0001$), while acidic amino acids prefer cytidine ($p < 0.0001$). Also, while these 'positive' contributions of amino acid–nucleotide interactions to protein–nucleic acid specificity are interesting, the 'negative' contributions, in which particular amino acids may avoid particular nucleotides (and vice versa), may also ultimately be important for determining specificity. The comprehensive access to amino acid–nucleotide interactions that is provided by AANT should greatly facilitate researchers attempting to understand, statistically or otherwise, the specificity of protein–nucleic acid interactions.

Even a basal analysis of the database provides an example of how the aggregated data can yield insights into individual protein–nucleic acid interactions. Proline is not normally known to be involved in nucleic acid interactions; indeed, prolines are the least utilized of any amino acid in the current 930 structures in AANT (Table 1). There has been one reported example of a RNA binding motif that contains proline residues; the herpesvirus U$_S$11 protein contains roughly 27 Arg-X-Pro repeats (22). However, in the context of this motif the proline residues help position the arginine side chains all along one face of a poly-L-proline II helix, and it is the 'arginine face' of this helix that actually makes contact with the nucleic acid (23,24).

The infrequent use of proline as an amino acid for contacting nucleic acids can also be readily observed in the 'AANTarctica' graph (Fig. 3), which also reveals interesting statistical anomalies. For instance, there are numerous proline–uridine interactions (29 occurrences) relative to proline–thymidine interactions (two occurrences, despite the fact that there are many more protein–DNA complexes in the PDB). The discrepancy between recognition of uridine and thymidine is also apparent for aspartate (94 occurrences in protein–RNA complexes, but only nine in protein–DNA complexes). The larger data set available through AANT makes these discrepancies more evident than in previous analyses. Luscombe and co-workers (9, their table 2) noted a single hydrogen-bond interaction between proline and DNA; AANT has captured eight such interactions thus far, while Treger and Westhof (10, their table 9) found 10 hydrogen-bond interactions between proline and RNA, compared with the 92 currently found in AANT. In addition, the breakdown of interaction data available via AANT makes apparent an interesting trend that was not noted in either of these two earlier studies with smaller data sets: proline contacts the nucleobase most often (75%) in protein–DNA interactions, but instead contacts the ribose most often (75%) in protein–RNA interactions.

Upon closer examination (an examination that was facilitated by the ability to extract lists of protein–nucleic acids from the 3D renderings of amino acid–nucleotide interactions), the

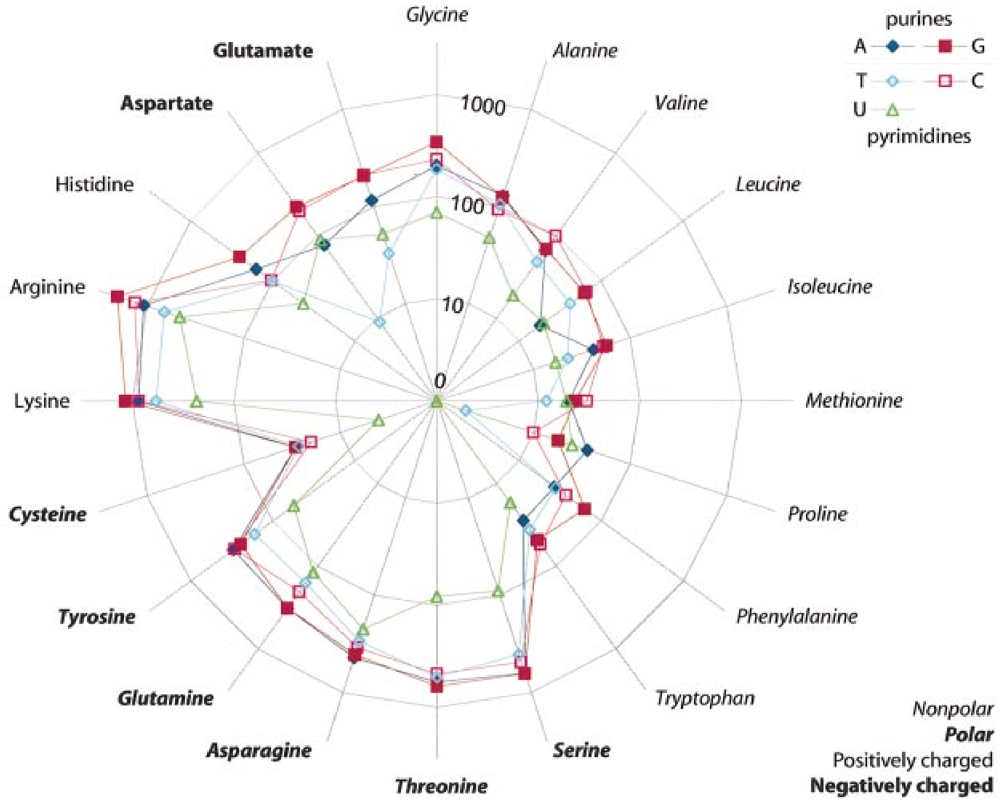


Figure 3. An 'AANTarctica' representation of the number of interactions between particular amino acids and particular nucleotides. Each interaction is defined by a marker at the intersection of one of 20 gray radial lines representing amino acids and one of five colored curves representing nucleotides. The distance of the marker from the center of the graph varies with the common logarithm of the number of interactions it represents. Nucleotides are depicted by the following markers: purines: A, solid dark blue diamond; G, solid dark maroon square; pyrimidines: C, hollow bright magenta square; T, hollow bright cyan diamond; U, hollow bright green triangle. Amino acids are labeled and organized into categories by label font: non-polar, medium italic; polar, bold italic; positively charged, medium roman; negatively charged, bold roman.

proline–uridine interactions appear to occur within one major protein type, that of ribosomal protein subunits. All of the current proline–uridine interactions in AANT involve either the S12 or S17 protein in the 30S ribosomal complex. Other ribosomal proteins also appeared to utilize proline to interact with rRNA. Many adenine:proline interactions can be found in ribosomal proteins, including L11, S2, S8, S10 and S11. Extending the examination to interactions with cytidine and guanosine again yields primarily contacts between ribosomal proteins and ribosomal RNA. The fact that the preponderance of proline contacts occurs in ribosomal proteins may merely be a function of the skewing of the database towards such proteins, or there may be a more interesting functional explanation. The proline backbone appears to interact closely with the ribose sugars of nucleotides. It is possible that these types of interactions are part of a strategy to promote the close approach of a protein over a large surface area, allowing the peptide backbone to extend (via proline 'kinks') along the surface of the ribosome.

In support of this notion, the other major sources of proline contacts, besides the ribosome, are other structures in which a protein and a RNA molecule contact one another over a large surface area, such as tRNA synthetases. Adenosine–proline contacts were seen in glutamyl-, aspartyl- and methionyl-tRNA synthetases, cytidine interactions were observed in aspartyl- and valyl-tRNA synthetases, and guanosine contacts in glutaminyl-tRNA synthetase. While these hypotheses are unproven at the moment, they provide an example of the type of insights that may be possible by using AANT as a tool for structure analysis.

AANT represents a unique tool that will allow structural biologists who study nucleic acid binding proteins or biochemists who study protein–nucleic acid interactions to better understand their particular molecule or interaction in the context of all possible protein–nucleic acid interactions. For example, while researchers commonly compare proteins that fall within a given structural class (e.g. helix–turn–helix proteins) it is difficult to find commonalities that do not rely upon a given structure (e.g. arginine-rich motifs) or that may extend across classes (25). Since AANT visually represents the geometries of individual amino acid–nucleotide interactions, common or new types of protein–nucleic acid interfaces can be quickly evaluated merely by scanning. Nascent structural hypotheses can then be validated or rejected by generating summaries of the proteins and nucleic acids involved, or by tabulating the relative frequencies of different types of interactions.

One of the major reasons for creating AANT was to better enable a variety of protein and nucleic acid engineering endeavors. First, in choosing which proteins or nucleic acids to engineer, the linear representations of protein–nucleic acid interactions that AANT can generate (Fig. 2) will assist researchers in identifying interfaces that are concentrated within short sequence stretches, and thus that can potentially be most easily engineered by cloning mutant oligonucleotides or PCR products. Second, the modular dissection of protein–nucleic acid complexes into amino acid–nucleotide complexes should facilitate the design of novel protein–nucleic acid interfaces. Previously, modular dissections of nucleic acid structures provided a database of examples of nucleotide–nucleotide interactions that allowed new nucleic acid structures to be accurately modeled *de novo* (26). A similar approach could presumably be taken using the amino acid–nucleotide interaction data provided by AANT. It is possible that the experimentally determined interactions could be modularly introduced into extant protein–nucleic acid scaffolds, replacing sterically similar interactions and altering specificities.

FUTURE DIRECTIONS

We are in the process of introducing several improvements to AANT. We intend to include predicted interactions other than hydrogen bonds, such as stacking interactions and van der Waals interactions, and will include these as separate interaction classes. We also plan to give users more options for choosing whether to display and consider interactions from redundant or alternative structures of the same protein–nucleic acid interaction. This should allow a user to more conveniently determine whether a statistic or correlation is due to skewing of the entries in the PDB. Finally, we propose to improve the utility of the database by adding a search algorithm that will identify sets of proteins that utilize similar amino acid geometries for interacting with their nucleic acid ligands.

ACKNOWLEDGEMENTS

This work was supported by Office of Naval Research Grant number N00014-99-1-0861.

REFERENCES

1. Weiss,M.A. and Narayana,N. (1998) RNA recognition by arginine-rich peptide motifs. *Biopolymers*, **48**, 167–180.
2. Reedstrom,R.J., Brown,M.P., Grillo,A., Roen,D. and Royer,C.A. (1997) Affinity and specificity of trp repressor–DNA interactions studied with fluorescent oligonucleotides. *J. Mol. Biol.*, **273**, 572–585.
3. Otwinowski,Z., Schevitz,R.W., Zhang,R.G., Lawson,C.L., Joachimiak,A., Marmorstein,R.Q., Luisi,B.F. and Sigler,P.B. (1988) Crystal structure of trp repressor/operator complex at atomic resolution. *Nature*, **335**, 321–329.
4. Draper,D.E. (1999) Themes in RNA–protein recognition. *J. Mol. Biol.*, **293**, 255–270.
5. Leulliot,N. and Varani,G. (2001) Current topics in RNA–protein recognition: control of specificity and biological function through induced fit and conformational capture. *Biochemistry*, **40**, 7947–7956.
6. Tao,J. and Frankel,A.D. (1992) Specific binding of arginine to TAR RNA. *Proc. Natl Acad. Sci. USA*, **89**, 2723–2736.
7. Jones,S., van Heyningen,P., Berman,H.M. and Thornton,J.M. (1999) Protein–DNA interactions: A structural analysis. *J. Mol. Biol.*, **287**, 877–896.
8. Jones,S., Daley, D.T., Luscombe,N.M., Berman,H.M. and Thornton,J.M. (2001) Protein–RNA interactions: a structural analysis. *Nucleic Acids Res.*, **29**, 943–954.
9. Luscombe,N.M., Laskowski,R.A. and Thornton,J.M. (2001) Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucleic Acids Res.*, **29**, 2860–2874.
10. Treger,M. and Westhof,E. (2001) Statistical analysis of atomic contacts at RNA–protein interfaces. *J. Mol. Recognit.*, **14**, 199–214.
11. Westbrook,J., Feng,Z., Chen,L., Yang,H. and Berman,H.M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, **31**, 489–491.
12. McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen-bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.
13. Pabo,C.O. and Nekludova,L. (2000) Geometric analysis and comparison of protein–DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol.*, **301**, 597–624.
14. Tou,J.T. and Gonzales,R.C. (1977) *Pattern Recognition Principles*. 2nd edn. Addison-Wesley, Reading, MA.
15. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.
16. Schwede,T., Kopp,J., Guex,N. and Peitsch,M.C. (2003) SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res.*, **31**, 3381–3385.
17. Suzuki,M. (1994) A framework for the DNA–protein recognition code of the probe helix in transcription factors: the chemical and stereochemical rules. *Structure*, **2**, 317–326.
18. Mandel-Gutfreund,Y., Schueler,O. and Margalit,H. (1995) Comprehensive analysis of hydrogen-bonds in regulatory protein–DNA complexes: in search of common principles. *J. Mol. Biol.*, **253**, 370–382.
19. Kono,H. and Sarai,A. (1999) Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*, **35**, 114–131.
20. Cheng,A.C., Chen,W.W., Fuhrmann,C.N. and Frankel,A.D. (2003) Recognition of nucleic acid bases and base-pairs by hydrogen-bonding to amino acid side-chains. *J. Mol. Biol.*, **327**, 781–796.
21. Seeman,N.C., Rosenberg,J.M. and Rich,A. (1976) Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl Acad. Sci. USA*, **73**, 804–808.
22. Khoo,D., Perez,C. and Mohr,I. (2002) Characterization of RNA determinants recognized by the arginine- and proline-rich region of Us11, a herpes simplex virus type 1-encoded double-stranded RNA binding protein that prevents PKR activation. *J. Virol.*, **76**, 11971–11981.
23. Roller,R.J., Monk,L.L., Stuart,D. and Roizman,B. (1996) Structure and function in the herpes simplex virus 1 RNA-binding protein U_s11: mapping of the domain required for ribosomal and nucleolar association and RNA binding *in vitro*. *J. Virol.*, **70**, 2842–2851.
24. Gresh,N. (1996) Can a polyproline II helical motif be used in the context of sequence-selective major groove recognition of B-DNA? A molecular modelling investigation. *J. Biomol. Struct. Dyn.*, **14**, 255–273.
25. Pabo,C.O. and Sauer,R.T. (1992) Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.*, **61**, 1053–1095.
26. Leclerc,F., Cedergren,R. and Ellington,A.D. (1994) A three-dimensional model of the Rev-binding element of HIV-1 derived from analyses of aptamers. *Nature Struct. Biol.*, **1**, 293–300.